

# Synthesis of Tongue Motion and Acoustics from Text using a Multimodal Articulatory Database

Ingmar Steiner<sup>1,2</sup>, Sébastien Le Maguer<sup>1</sup>, and Alexander Hewer<sup>1,2,3</sup>

<sup>1</sup>*Cluster of Excellence “Multimodal Computing and Interaction”,  
Saarland University, Saarbrücken, Germany*

<sup>2</sup>*German Research Center for Artificial Intelligence  
(DFKI GmbH), Saarbrücken, Germany*

<sup>3</sup>*Saarbrücken Graduate School in Computer Science, Saarbrücken,  
Germany*

## Abstract

We present an end-to-end text-to-speech (TTS) synthesis system that generates audio and synchronized tongue motion directly from text. This is achieved by adapting a 3D model of the tongue surface to an articulatory dataset and training a statistical parametric speech synthesis system directly on the tongue model parameter weights. We evaluate the model at every step by comparing the spatial coordinates of predicted articulatory movements against the reference data. The results are encouraging, and our approach can be adapted to add an articulatory modality to conventional TTS applications without the need for extra data.

## 1 Introduction

The sound of human speech is the direct result of production mechanisms in the human vocal tract. Air flows from the lungs through the glottis, whose vocal folds can be set to vibrate, the sound of which is then filtered by the shape of the tongue, lips, and other articulators, generating what we perceive as audible signals such as spoken language. Researchers in phonetics and linguistics have studied these speech production mechanisms for many years, but while the acoustic signal and facial movements can be observed and measured directly, doing the same for partially or fully hidden articulators such as the tongue and glottis is not as straightforward.

Consequently, sensing and imaging techniques have been applied to the challenge of observing speech production mechanisms *in vivo*, which has greatly improved our understanding of these processes. The corresponding modalities include, fluoroscopy [25], ultrasound tongue imaging (UTI) [35], X-ray microbeam (XRMB) [38], electromagnetic articulography (EMA) [12, 32], and real-time magnetic resonance imaging (MRI) [26, 27], among others. Some of these involve health hazards (due to ionizing radiation), and all are more or less invasive, but they produce *biosignals* which, in combination with simultaneous acoustic

recordings, represent multimodal articulatory speech data. The benefits are tempered by the challenges of processing the imaging and/or point-tracking data, which in the field of speech processing has created new opportunities for collaboration with areas such as medical imaging and computer vision.

The biosignals that can be obtained using such modalities to record spoken language, provide opportunities to greatly enhance models of speech by integrating measurements of the underlying processes directly with the acoustic signal. This leads to more elegant and powerful approaches to speech analysis and synthesis [19, 21, 30]. However, it must be borne in mind that all of the biosignals produced by the modalities mentioned above represent a sampling of the articulators that is *sparse* in the temporal domain, the spatial domain, or both.

Depending on the manner in which the data is used for analysis or applications, the resolution may need to be increased, but the missing samples cannot be restored without prior knowledge, typically provided by a statistical model trained on other data.

In this study, we present an approach to multimodal text-to-speech (TTS) synthesis that generates the fully animated, three-dimensional (3D) surface of the tongue, synchronized with synthetic audio, using data from a single-speaker, articulatory corpus that includes EMA motion capture of three tongue flesh-points [29]. The audio and articulatory motion are synthesized using the hidden Markov model (HMM) based synthesis (HTS) framework [44], while the surface restoration is performed by means of a multilinear statistical tongue model [11] trained on a multi-speaker, volumetric MRI dataset [28]. The potential application domains of this approach include audiovisual speech synthesis, or computer-assisted pronunciation training (CAPT).

## 1.1 Background

Deriving models suitable for producing speech related tongue motion is an active field of research. Such models can, for example, help to analyze and understand articulatory data that is very sparse in the spatial domain. Ideally, such tongue models should offer a good compromise between accuracy of the generated shape and the available degrees of freedom (DoF) for manipulating it. This means that so-called biomechanical models such as those presented by Lloyd et al. [23], Xu et al. [40], Wrench and Balch [39], or Yu et al. [42] might be too complex for this purpose. These models aim at simulating the underlying mechanics of the human tongue as closely as possible, and can be used to visualize existing articulatory data.

Geometric tongue models are less complex than their biomechanical counterparts. Here, we distinguish between generic and statistical tongue models. Generic tongue models are 3D models of the tongue that may be deformed and animated by using standard methods in computer graphics.

Statistical tongue models, on the other hand, are constructed by analyzing the DoFs of the tongue shape in recorded articulatory data, like for example MRI recordings of speech related vocal tract shapes. Roughly speaking, such an analysis can be carried out in two ways. The first variant investigates shape variations related to the tongue pose that are specific for speech production. Examples of such approaches are the works by Engwall [7] and Badin and Serrurier [3], Badin et al. [4] who examined those variations in 3D MRI scans from a single

speaker, respectively. These methods only estimate the DoF that are tongue pose related, while shape variations that may describe anatomical differences are missing.

Another class of methods aims at investigating those anatomy and tongue pose related shape variations separately. This paradigm offers the following advantages: First, the results give access to tongue models that may be adapted to new speakers. Second, this type of analysis may also provide insight into how anatomical differences affect human articulation. For two-dimensional (2D) MRI, such work was conducted, e.g., by Hoole et al. [13] and Ananthakrishnan et al. [1]. Zheng et al. [45] investigated those variations in a sparse point cloud extracted from 3D MRI. Most recently, we performed such an analysis on mesh representations of the tongue that were extracted from 3D MRI scans [11].

Such geometrical models have been successfully used in previous work to generate animations from provided articulatory data: Katz et al. [15] presented a real-time visual feedback system that deforms a generic tongue model by using EMA data. However, due to the generic model, their approach did not take anatomical differences into account. A statistical model was used in the approach by Badin et al. [5]. They used the data of one speaker to derive the tongue model and used the EMA data of the same speaker to animate it. Engwall [8] followed a similar approach. Our own previous work utilized a multilinear statistical model to visualize EMA data, which allowed it to be adapted to different speakers [14].

Independently, there is a growing body of work on application-oriented research to combine articulatory data, and features derived from it, with speech technology applications, such as to recover articulatory movements from the acoustic signal [“articulatory inversion mapping”, cf. 17, 24, for example], provide articulatory control for reactive TTS synthesis [e.g., 2, 22], or predict sparse articulatory movements from a symbolic representation [e.g., 6, 21]. However, to our knowledge, no previous study has succeeded in presenting an end-to-end system to directly generate the motion of a full 3D model of the tongue surface from text, particularly one that can be easily adapted to the anatomy of different speakers.

## 2 Method

### 2.1 Multilinear Shape Space Model

Given a number of 3D points on the tongue surface, we can fit a statistical model of the entire tongue to those points. This model controls the shape of a triangular mesh formed by 3D vertices.

The tongue model used for this study was created from a database of volumetric MRI scans of the vocal tracts of 12 speakers, each of whom produced the same set of English vowels and sustained consonants. The volumetric data was processed automatically by applying a diffusion method for denoising, followed by an image segmentation technique. Then a template mesh was deformed to fit the resulting 3D point clouds; this approach is described in detail in [10].

By training a statistical model on the resulting set of mesh deformations, we obtain a multilinear shape space model of the tongue that is able to separate speaker identity (i.e., vocal tract anatomy) from phoneme identity (i.e., the

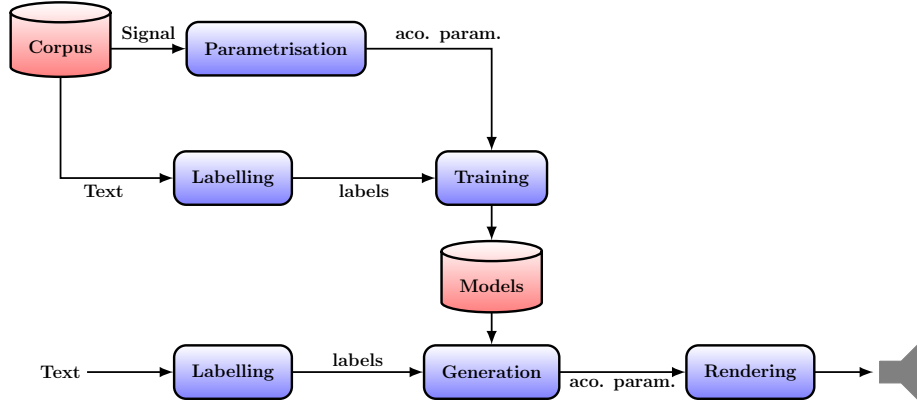


Figure 1: Architecture of the HTS system as described in [43]

tongue shape for a given sound). Based on the training data, this multilinear tongue model contains 12 speaker parameters, and 13 phoneme parameters, of decreasing significance within each corresponding parameter subspace.

The multilinear tongue model can be fit to new data, in the form of 3D point clouds, with very low latency. In fact, the target data can be so sparse that only a few points are sufficient to adapt the model, so instead of segmented MRI data, it can also be fit to the tongue surface fleshpoints represented by EMA data. Moreover, we can avoid generating unrealistic tongue shapes during this optimization because the model is able to evaluate the probability of a tongue shape. When fitting to a stream of EMA data in this way, the identity of the speaker does not change, and so the multilinear model can be adapted and fixed in the speaker subspace, yielding the equivalent of a principal component analysis (PCA) shape space model for that speaker.

With a fixed coordinate in the speaker subspace, the tongue model varies only in the 13-dimensional phoneme subspace; i.e., it is reduced to 13 DoFs. Moreover, because the significance of parameters is inversely correlated with their order, the phoneme subspace can also be clipped to fewer parameters, producing only marginally less accurate deformations. A detailed evaluation of the multilinear model under such conditions is provided in [11].

## 2.2 Multimodal Statistical Parametric Speech Synthesis

The HMM based synthesis (HTS) framework [43] is a standard statistical parametric speech synthesis system. The architecture, illustrated in Figure 1, is decomposed into four main parts: the parametrization of the signal, the training of the models, the parameter generation, and the signal rendering.

The focus of our study impacts the parametrization and the rendering stage. Therefore, we use the standard training stage (described in [43]) and the standard parameter generation algorithms (described [37]).

The parametrization of the signal can be performed using any suitable signal processing tool. The important part is to keep the parametrization and the signal rendering consistent. In the standard procedure, this is generally accom-

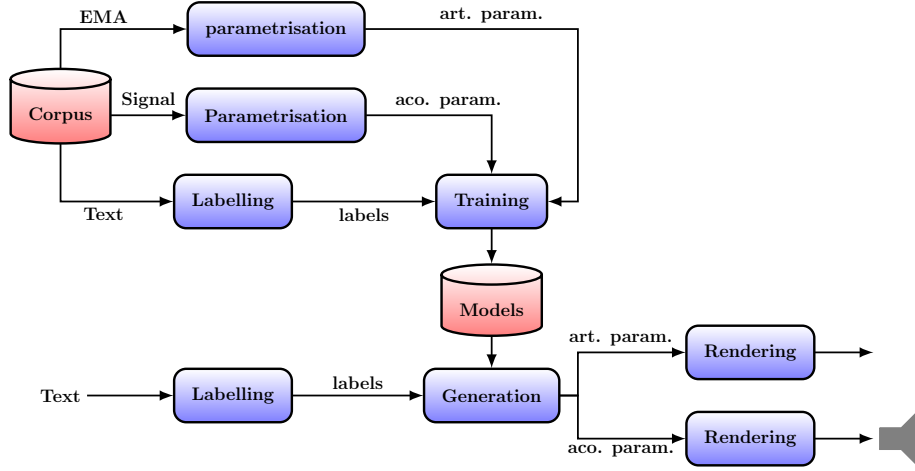


Figure 2: Adapted architecture for multimodal HMM based synthesis

plished by coupling STRAIGHT [16] with a mel log spectrum approximation (MLSA) filter [9]. First, STRAIGHT is used to extract the spectrum, the fundamental frequency ( $F_0$ ), and the aperiodicity. Generally, the  $F_0$  values are transformed into the logarithmic domain, to be more consistent with human hearing. Furthermore, the order of the spectrum and the aperiodicity is too high. Therefore, the MLSA filter is used to parametrize these coefficients and to obtain the mel-generalized cepstral coefficients (MGC) coefficients and the aperiodicity per band (BAP), respectively.

In this study, we propose to not only consider the parametrization of the acoustic signal but also the parametrization of speech articulation. In previous studies [19, 20, 21], EMA data was used as the articulatory representation. In the present study, we work towards replacing the EMA data by the tongue model parameters. Therefore, our goal is to train on the trajectories of the tongue model parameters using HTS. Figure 2 presents the details of the modified architecture.

## 3 Experiments

### 3.1 Database

The data used for the experiments in this study is taken from the *mngu0* corpus, specifically the “day 1” EMA subset [29], which contains acoustic recordings, time-aligned phonetic transcriptions, and EMA motion capture data (sampled at 200 Hz using a Carstens AG500 articulograph).<sup>1</sup> We selected the “basic”

<sup>1</sup>From the *mngu0* website, <http://mngu0.org>, we downloaded the following distribution packages:

- Day1 basic audio data downsampled to 16 kHz (v1.1.0)
- Day1 basic EMA data, head corrected and unnormalized (v1.1.0)
- Day1 transcriptions, Festival utterances and ESPS label files (v1.1.1)

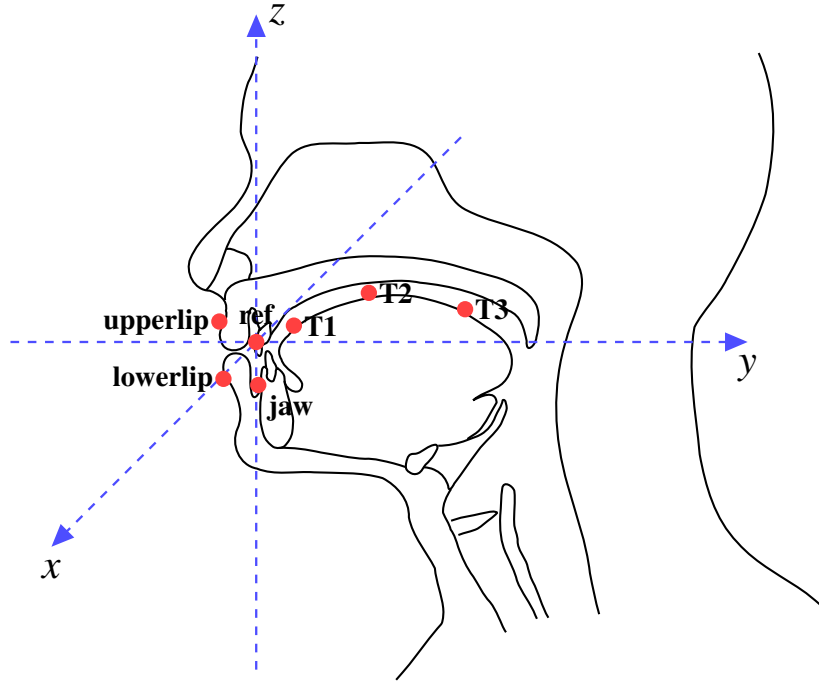


Figure 3: EMA coil layout in the “day 1” subset of the *mngu0* corpus. All coils are close to the mid-sagittal plane. The **ref** coil on the upper incisors forms the origin of the coordinate space; the horizontal and vertical axes represent the  $y$  and  $z$  dimensions in the data, respectively, while the  $x$  axis is perpendicular to the image plane. Adapted from [29].

Table 1: EMA coil labels and locations in the “day 1” subset of the *mngu0* corpus.

label	location
<b>T1</b>	tongue tip
<b>T2</b>	tongue body
<b>T3</b>	tongue dorsum
<b>upperlip</b>	upper lip
<b>lowerlip</b>	lower lip
<b>ref</b>	upper incisor
<b>jaw</b>	lower incisor

Table 2: Global evaluation measures for the acoustic synthesis baseline.

id	mean	std. dev.	conf. int.
<b>rms dur</b>	42.00	18.29	2.93
<b>rms f0</b>	188.52	76.92	12.33
<b>vuvrate</b>	12.03	3.94	0.63
<b>mcdist</b>	2.45	0.22	0.04

(as opposed to the “normalized”) release variant of the EMA data, because it preserves the silent (i.e., non-speech) intervals, as well as the 3D nature and true spatial coordinates of the sensor data (after head motion compensation). The EMA coil layout for this data is shown in Figure 3; the coils are explained in Table 1.

In order to manipulate the EMA data more flexibly, the files were first converted from the binary Edinburgh Speech Tools (EST) format to a JSON structure. Invalid values (i.e., NaN) were replaced by linear interpolation. No further modification, in particular no smoothing, was applied.

From the provided acoustic data, signal parameters were extracted using STRAIGHT [16] with a frame rate of 200 Hz, matching that of the EMA data. As we follow the standard HTS methodology, we also kept the same parameters. Therefore, our signal parameters are 50 MGC, 25 BAP and one coefficient for the  $F_0$ .

From the 1354 utterances in the data, 152 (11.20 %) were randomly selected and held back as a test set; the remaining 1202 utterances were used as the training set to build HTS synthesis voices.

### 3.2 Acoustic Synthesis

As a baseline, we first built a conventional TTS system using the acoustic data only. This served mainly to validate our voicebuilding process and ensure that the transcriptions provided, and labels generated from them, along with the acoustic signal parameters, were able to generate audio of sufficient quality. Accordingly, we did not conduct a formal subjective listening test, instead evaluating this baseline experiment using objective measures.

We synthesized the 152 utterances in the test set, imposing the acoustic phone durations from the provided transcriptions to allow direct comparison with the natural recordings. The objective evaluation we conducted relies on widely used evaluation metrics. For the duration evaluation, we calculated the root mean square (RMS) duration at the phone level in ms. For the  $F_0$ , we use two measures: the voiced-unvoiced (VUV) error rate percentage to check the prediction of the  $F_0$ , and the RMS in cent. The latter measure focuses on the frames which are voiced in both conditions (original and predicted  $F_0$ ). Furthermore, it is a log scale measure adapted to the human perception. Finally, for the spectrum part, we computed a mel cepstral distortion (MCD) in dB. Except for the duration, all parameters were evaluated at a frame level. Based on these measures, we can compare our results to previous studies, such as the one presented by Yokomizo et al. [41].

The results of this evaluation are given in Table 2 and comprise the mean,

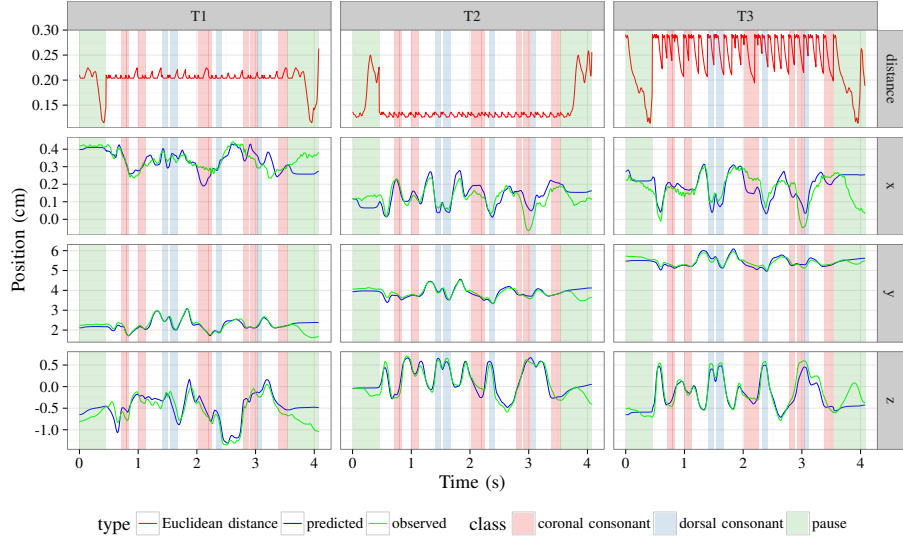


Figure 4: Observed and predicted position trajectories (along the  $x$ ,  $y$ , and  $z$  axis), and Euclidean distance (top), for the tongue EMA coils (T1, T2, T3) for one test utterance, using combined acoustic and EMA synthesis. The utterance is “Because these deer are gregarious, they go about in groups”. Based on the provided transcriptions, intervals containing silent (pause) and coronal and dorsal consonants have been highlighted.

standard deviation, and confidence interval with a  $p$  value at 5%. Compared to [41], we achieved slightly better results, notwithstanding the different dataset. Therefore, we can conclude that our acoustic prediction is consistent with the state of the art in HTS.

### 3.3 Combined Acoustic and EMA Synthesis

In an early multimodal fusion paradigm, we combined the acoustic signal parameters with the 3D positions of the seven EMA coils shown in Table 1, increasing the vector size by 21, to 97 parameters per frame. Using the HTS framework, we then built another TTS system from this multimodal data.

Synthesizing the test set in this way, we obtained, in addition to the audio, synthetic trajectories of predicted EMA coil positions. To evaluate the combined acoustic and EMA synthesis, we computed the same objective measures as in Section 3.2. We also computed the Euclidean distance in space between the observed and predicted positions for the EMA coils. The results of this evaluation are given in Table 3. We see that the differences in the acoustic measures compared to the acoustic-only synthesis (cf. Table 2) are negligible.

The comparison between the observed and predicted trajectories for one test utterance is illustrated in Figure 4. The observed and predicted (synthesized) positions of the three tongue coils are shown in each of the three dimensions in the data, along with the Euclidean distance. Silent intervals and those classified



Table 3: Global evaluation for the combined acoustic and EMA synthesis.

id	mean	std. dev.	conf. int.
<b>rms dur</b>	41.93	19.04	3.05
<b>rms f0</b>	188.43	63.70	10.21
<b>vuvrate</b>	12.14	3.84	0.62
<b>mcdist</b>	2.45	0.23	0.04
<b>euclidist T3</b>	0.21	0.15	$8.57 \times 10^{-4}$
<b>euclidist T2</b>	0.21	0.15	$9.00 \times 10^{-4}$
<b>euclidist T1</b>	0.22	0.16	$9.44 \times 10^{-4}$
<b>euclidist ref</b>	0.02	0.01	$6.97 \times 10^{-5}$
<b>euclidist jaw</b>	0.13	0.07	$3.80 \times 10^{-4}$
<b>euclidist ulip</b>	0.07	0.04	$2.21 \times 10^{-4}$
<b>euclidist llip</b>	0.14	0.09	$5.45 \times 10^{-4}$

as coronal and dorsal consonants ([t, d, n, l, s, z, ʃ, ʒ, θ, ð] and [g, k, ŋ], respectively), based on the provided phonetic transcription, have been highlighted. This helps visualize the correspondence between gestures of the tongue tip (coil T1) and tongue back (coils T2 and T3) for coronal and dorsal consonants, respectively, and the phonetic units they produce.

Several points merit discussion.

First of all, there are large mismatches between the observed and predicted tongue EMA coil positions during the silent (pause) intervals at the beginning and end of the utterance. This can be attributed to the fact that the wide range of the speaker’s tongue movements during non-speech intervals was not distinguished in the annotations, but invariably annotated with the same pause label. However, there are at least two very distinct shapes for the tongue during such silent intervals, including a “rest” and a “ready” position (just before speech is produced), in addition to other complex movements such as swallowing. In the absence of distinct labels corresponding to these positions and movements, none of this silent variation can be captured by the HMMs trained on this data; instead, the tongue coils are unsurprisingly predicted to hover around global means.

Secondly, there is noticeable oversmoothing, and target extrema are not quite reached. This is typical, and can be attributed to, the HMM based synthesis technique, despite the integration of global variance. The dynamics, however, are well represented, and the predicted positional trajectories, as well as their derivatives, match the observed reference quite closely.

The  $x$  axis appears to suffer from a greater amount of prediction error than the  $y$  or  $z$  axes. However, it should be noted that the positional variation along the  $x$  axis is an order of magnitude smaller than that along the  $y$  axis. It must also be borne in mind that nearly all of the speech-related movements occur along the mid-sagittal plane, represented by the  $y$  (anterior/posterior)  $z$  (inferior/superior) axes; variation along the  $x$  axis corresponds to lateral movements, which are infrequent during speech.<sup>2</sup> Having said that, the  $x$  axis can

<sup>2</sup>Following this rationale, the “normalized” release variant of the *mngu0* EMA dataset actually consists of flattened, 2D data, with all coil positions projected onto the mid-sagittal plane.

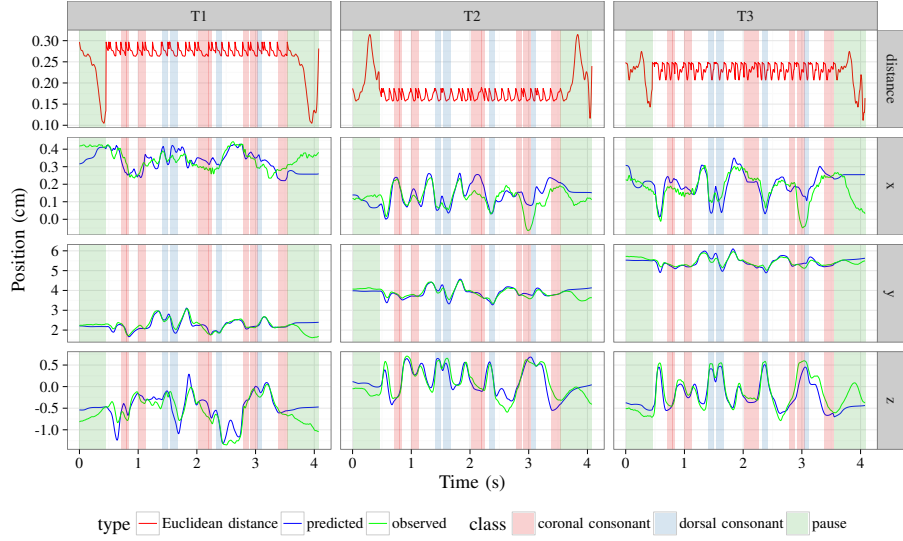


Figure 5: One test utterance produced using EMA-only synthesis; all other details are the same as in Figure 4.

serve to illustrate the physical coil locations on the tongue in the “day 1” recording session; to wit, the tongue tip coil is actually attached out of plane, a few millimeters to one side.

The Euclidean distances *during* speech are in the sub-millimeter range, indicating that the predictions of EMA coil positions in space are accurate to within the precision of the EMA measurements themselves. However, there appears to be a certain amount of fluctuation with a more or less regular range and shape. The peaks of this fluctuation appear to correlate with spikes in the RMS channels of the provided EMA data, which supports the hypotheses that it is either an artifact of the algorithm which calculates the coil positions and orientations from the raw amplitudes [34], or measurement noise in the articulograph itself [18], or, conceivably, a combination of both factors. Of course, the noise in the Euclidean distance analysis is a direct consequence of our decision to refrain from smoothing the provided EMA data.<sup>3</sup>

### 3.4 EMA Synthesis

While the combined acoustic and EMA synthesis produced satisfactory results, the requirement to train the system on a multimodal dataset such as *mngu0* represents a significant drawback; compared to the reasonably wide availability of conventional, acoustic databases designed for speech synthesis, the number of suitable articulatory databases is extremely low. Encouraged by the practical equivalence in the evaluation of the acoustic measures described in Section 3.2 and Section 3.3, we therefore considered the question of decoupling the EMA

<sup>3</sup>Perhaps the RMS jitter in the unsmoothed measurements could also be exploited in adaptive EMA denoising.

Table 4: Global evaluation for the EMA-only synthesis.

id	mean	std. dev.	conf. int.
<b>rms dur</b>	53.73	20.74	3.32
<b>eucdist T3</b>	0.22	0.14	$8.32 \times 10^{-4}$
<b>eucdist T2</b>	0.22	0.15	$9.01 \times 10^{-4}$
<b>eucdist T1</b>	0.23	0.16	$9.44 \times 10^{-4}$
<b>eucdist ref</b>	0.02	0.01	$6.80 \times 10^{-5}$
<b>eucdist jaw</b>	0.13	0.07	$3.87 \times 10^{-4}$
<b>eucdist ulip</b>	0.07	0.04	$2.19 \times 10^{-4}$
<b>eucdist llip</b>	0.15	0.09	$5.36 \times 10^{-4}$

Table 5: Global evaluation for the EMA-only synthesis restricted to the tongue coils.

id	mean	std. dev.	conf. int.
<b>rms dur</b>	61.20	21.64	3.47
<b>eucdist T3</b>	0.22	0.14	$8.46 \times 10^{-4}$
<b>eucdist T2</b>	0.22	0.15	$8.76 \times 10^{-4}$
<b>eucdist T1</b>	0.23	0.16	$9.12 \times 10^{-4}$
<b>eucdist ref</b>	0.02	0.01	$7.02 \times 10^{-5}$

synthesis completely from the acoustic data. Accordingly, we used the HTS framework to build another TTS system trained only on the EMA data, without the acoustic parameters.

The evaluation of the RMS duration and Euclidean distances between the predicted and observed EMA coils is given in Table 4. As we can see, the results are nearly identical to those in Table 3, which confirms the validity of this approach. Figure 5 visualizes the comparison between the observed and predicted trajectories for one test utterance.

### 3.5 Tongue-only EMA Synthesis

In order to focus on the tongue in the following section, we first needed to investigate how far the tongue coil EMA positions can be predicted in isolation from the remaining EMA coils. To this end, we created a modified version of the TTS system described in the previous section, by including *only* the tongue coils (T1, T2, and T3), and excluding the rest of the EMA data from the training set.

Table 5 gives the evaluation of the EMA synthesis restricted to the three tongue coils. Comparing these results with those in Table 4, we observe that apart from a slightly higher RMS duration, the values are virtually identical, which confirms the validity of this approach. As before, the comparison between the observed and predicted trajectories for one test utterance is shown in Figure 6. It should be noted that despite the removal of the EMA coil on the lower incisor, some residual jaw motion is implicitly retained in the movements of the tongue coils.

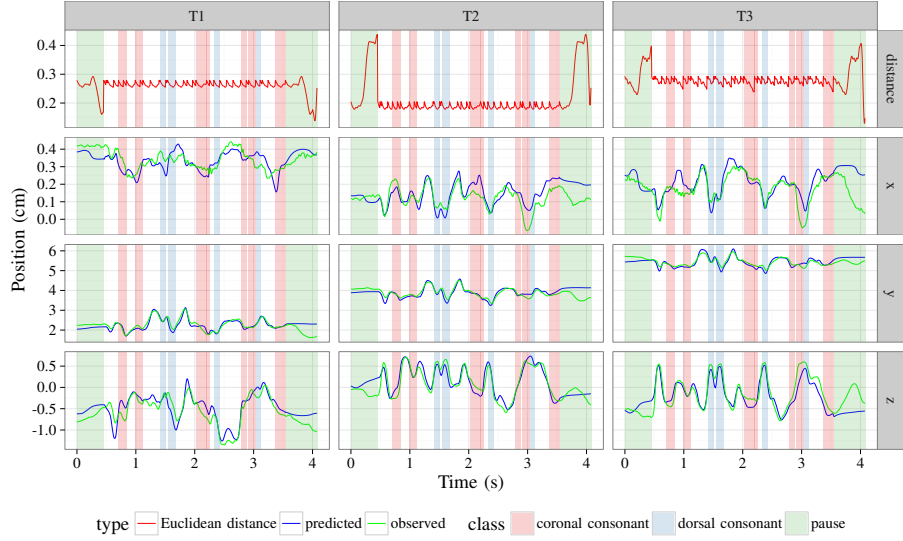


Figure 6: One test utterance produced using EMA-only synthesis restricted to the tongue coils; all other details are the same as in Figure 4.

Table 6: Global evaluation for the tongue model parameter weights synthesis.

id	mean	std. dev.	conf. int.
<b>rms dur</b>	77.66	26.51	4.25
<b>eucdist T3</b>	0.26	0.16	$9.43 \times 10^{-4}$
<b>eucdist T2</b>	0.28	0.17	0.00
<b>eucdist T1</b>	0.29	0.19	0.00
<b>eucdist ref</b>	0.05	0.01	$6.70 \times 10^{-5}$

### 3.6 Model-based Tongue Motion Synthesis

At last, having verified that the HTS framework can be used to synthesize audio and predict the movements of three tongue EMA coils using *separate* models trained on the *mngu0* database, we prepared a new kind of TTS system to predict the shape and motion of the entire tongue surface, by integrating a geometric 3D model into the process.

First, our multilinear shape space model of the tongue (cf. Section 2.1) was adapted to the identity of the speaker in the *mngu0* dataset in several steps.

1. We automatically selected three vertices on the template mesh that corresponded most closely with the three tongue coils in the EMA data.
2. For each frame of each utterance in the dataset, the multilinear tongue model was fitted to the three tongue EMA coil positions, and the values for the resulting parameter weights were stored.
3. Given the fitted model deformations, the average weights for the speaker

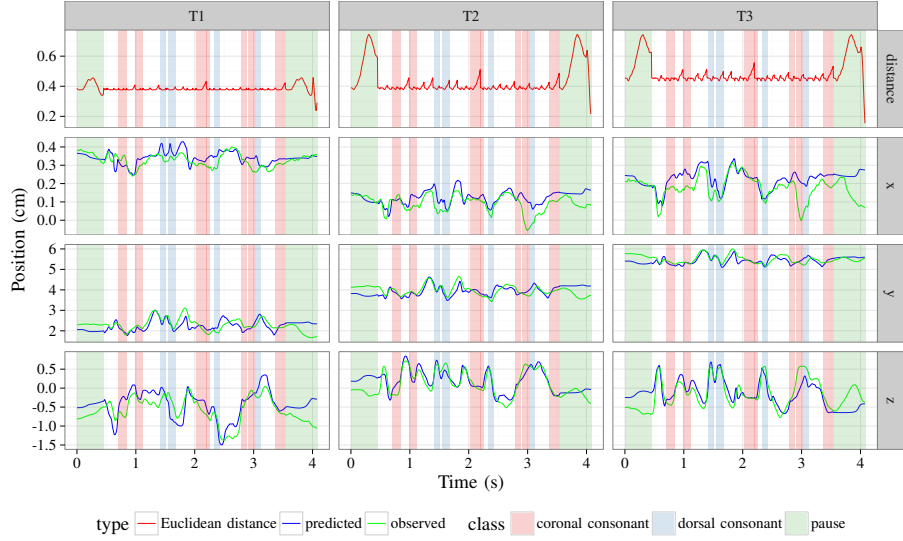


Figure 7: One test utterance produced using the tongue model parameter weights synthesis; all other details are the same as in Figure 4.

subspace were computed. The speaker subspace weights converge on their respective means after a few dozen utterances. If we then fix the speaker weights to these means, the model effectively becomes a single-speaker, PCA model.

4. The speaker-adapted tongue model was fitted to the three tongue EMA coil positions in the phoneme subspace for each frame of each utterance in the dataset, storing the resulting parameter weight values.

The trajectories of the fitted parameter weights in the phoneme subspace were taken as the training data, and we used the HTS framework to build for a new TTS system that predicts the tongue model parameter values directly from the input text. With a combination of this system and the conventional one described in Section 3.2, it is possible to first synthesize the acoustic speech signal, and to provide the predicted acoustic durations to guide the synthesis of corresponding tongue model parameters, which are then used to animate the 3D tongue model in real time.

To evaluate the performance of this system against the reference EMA data, we extracted the spatial coordinates of the vertices assigned during the adaptation step (see above) to produce synthetic trajectories that served as a virtual surrogate for predicted EMA data.

We evaluated this synthetic EMA data against the reference as before; Table 6 provides the RMS duration and Euclidean distances between the predicted and observed EMA coils, and one test utterance is visualized in Figure 7. It should be noted that the tongue model itself contains a temporal smoothing term, which ensures that a noisy sequence of input frames does not cause the 3D mesh to change shape or position too rapidly; this extra smoothing contributes

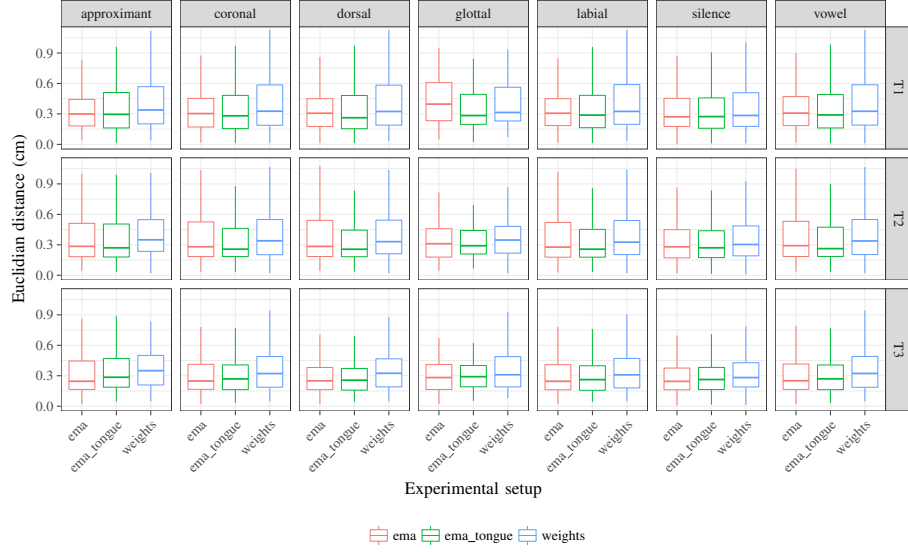


Figure 8: Distributions of Euclidean distances between observed and predicted tongue EMA coil positions for each experimental TTS setup (“ema”: EMA synthesis, “ema\_tongue”: EMA synthesis restricted to tongue coils, and “weights”: synthesis of tongue model parameter weights), split by phone class and tongue EMA coil.

to widespread target undershoot in the comparison. Overall, the results of this evaluation are very promising, and we can confirm that as far as possible with only three surface points on the tongue, the animation of the full tongue appears to closely match the observed reference.

To conclude this section, we have compared the distribution of Euclidean distances to the observed reference data over the entire test set, with all three experimental TTS systems (trained without acoustic data), and the results can be seen in Figure 8. The distances increase slightly as the non-tongue coils are excluded, and again as the direct EMA prediction is replaced by the synthesis of tongue model parameter weights. However, overall, the distances remain in the same range, which indicates that the approach presented here does not perform worse than direct synthesis of EMA data – while adding the full 3D tongue surface into the synthesis process.

## 4 Conclusion

In this study, we have presented a new process of synthesizing acoustic speech and synchronized animation of a full 3D surface model of the tongue. We used the HTS framework with a single-speaker, multimodal articulatory database containing EMA motion capture data. First, we demonstrated a conventional, fused multimodal approach, then separated the two modalities while ensuring that the objective evaluation measures remained comparable. Finally, we adapted a multi-linear statistical model of the tongue and integrated it into

the TTS process, and evaluated its accuracy by comparing the spatial coordinates of vertices on the model surface to the reference EMA data from the original speaker’s tongue movements. The results are very encouraging, and we believe that this will enable multimodal TTS applications that provide tongue animation with human-like performance.

It should be noted that the acoustic synthesis and predicted phone durations need not come from the same corpus as the one used for training the tongue model parameter synthesis system. Within certain limits, it would be straightforward to use a different, conventional TTS system with speech recordings from a different speaker in combination with this tongue model parameter synthesis, perhaps adapting it in the speaker subspace automatically or by hand, to generate a multimodal TTS application with plausible, speech-synchronized tongue motion, without the requirement of having articulatory data available for the target speaker.

However, there is clearly more work to be done, and in future research, we intend to refine and improve our system, and to evaluate it using human subjects who will rate it perceptually. Such a study can include intelligibility, such as the contribution of visible tongue movements during degraded, noisy, or absent audible speech. However, we also plan to assess the impact on perceived naturalness by integrating the tongue model into a realistic talking avatar [e.g., 31, 36], and investigating the importance of naturalistic tongue movements for the overall impression of such avatars in multimodal spoken interaction scenarios with artificial characters. This may also lead us to model distinct non-speech poses for the tongue, such as separate “rest” and “ready” positions.

Regarding the tongue model integration, we plan to further investigate such factors as the impact of reducing the number of parameters on synthesis performance, optimizing the vertex correspondence with EMA data, and exploring speaker adaptation using volumetric data, such as the MRI subset of the *mngu0* corpus [33].

## Acknowledgment

We are grateful to Korin Richmond, Phil Hoole, and Simon King for creating and releasing the *mngu0* database. Studies such as the one described in this paper would not be possible without such high-quality, open databases.

## References

- [1] Gopal Ananthakrishnan, Pierre Badin, Julián Andrés Valdés Vargas, and Olov Engwall. Predicting unseen articulations from multi-speaker articulatory models. In *Interspeech*, pages 1588–1591, September 2010. URL [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_1588.html](http://www.isca-speech.org/archive/interspeech_2010/i10_1588.html).
- [2] Maria Astrinaki, Alexis Moinet, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, and Thierry Dutoit. Mage – reactive articulatory feature control of HMM-based parametric speech synthesis. In *8th ISCA Workshop on Speech Synthesis (SSW)*, pages 207–211, August – September 2013. URL [http://ssw8.talp.cat/papers/ssw8\\_DS-2\\_Astrinaki.pdf](http://ssw8.talp.cat/papers/ssw8_DS-2_Astrinaki.pdf).

- [3] Pierre Badin and Antoine Serrurier. Three-dimensional linear modeling of tongue: Articulatory data and models. In *7th International Seminar on Speech Production (ISSP)*, pages 395–402, December 2006.
- [4] Pierre Badin, Gerard Bailly, Lionel Reveret, Monica Baciú, Christoph Segebarth, and Christophe Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553, July 2002. doi:10.1006/jpho.2002.0166.
- [5] Pierre Badin, Frédéric Elisei, Gérard Bailly, and Yuliya Tarabalka. An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker’s articulatory data. In *Articulated Motion and Deformable Objects*, pages 132–143. Springer, 2008. doi:10.1007/978-3-540-70517-8\_14.
- [6] Ming-Qi Cai, Zhen-Hua Ling, and Li-Rong Dai. Statistical parametric speech synthesis using a hidden trajectory model. *Speech Communication*, 72:149–159, September 2015. doi:10.1016/j.specom.2015.05.008.
- [7] Olov Engwall. A 3D tongue model based on MRI data. In *6th International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, October 2000. URL [http://www.isca-speech.org/archive/icslp\\_2000/i00\\_3901.html](http://www.isca-speech.org/archive/icslp_2000/i00_3901.html).
- [8] Olov Engwall. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41(2-3):303–329, October 2003. doi:10.1016/S0167-6393(02)00132-2.
- [9] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptative algorithm for mel-cepstral analysis of speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140, March 1992. doi:10.1109/ICASSP.1992.225953.
- [10] Alexander Hewer, Stefanie Wuhler, Ingmar Steiner, and Korin Richmond. Tongue mesh extraction from 3D MRI data of the human vocal tract. In *Perspectives in Shape Analysis*, pages 345–365. Springer, 2016. doi:10.1007/978-3-319-24726-7\_16.
- [11] Alexander Hewer, Stefanie Wuhler, Ingmar Steiner, and Korin Richmond. A multilinear tongue model derived from speech related MRI data of the human vocal tract. *arXiv preprint arXiv:1612.05005*, 2016. URL <https://arxiv.org/abs/1612.05005>.
- [12] Phil Hoole and Andreas Zierdt. Five-dimensional articulography. In *Speech Motor Control: New Developments in Basic and Applied Research*, pages 331–349. Oxford University Press, 2010.
- [13] Phil Hoole, Axel Wismüller, Gerda Leinsinger, Christian Kroos, Anja Geumann, and Michiko Inoue. Analysis of tongue configuration in multi-speaker, multi-volume MRI data. *5th Seminar on Speech Production*, pages 157–160, May 2000.



- [14] Kristy James, Alexander Hewer, Ingmar Steiner, and Stefanie Wuhler. A real-time framework for visual feedback of articulatory data using statistical shape models. In *Interspeech*, pages 1569–1570, September 2016. URL [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/2019.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/2019.html).
- [15] William Katz, Thomas F Campbell, Jun Wang, Eric Farrar, J Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, and Rob Rennaker. Opti-Speech: a real-time, 3D visual feedback system for speech training. In *Interspeech*, pages 1174–1178, September 2014. URL [http://www.isca-speech.org/archive/interspeech\\_2014/i14\\_1174.html](http://www.isca-speech.org/archive/interspeech_2014/i14_1174.html).
- [16] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4): 187–207, April 1999. doi:10.1016/S0167-6393(98)00085-5.
- [17] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester. Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, 121(2): 723, January 2007. doi:10.1121/1.2404622.
- [18] Christian Kroos. Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). *Journal of Phonetics*, 40(3):453–465, May 2012. doi:10.1016/j.wocn.2012.03.002.
- [19] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1171–1185, August 2009. doi:10.1109/tasl.2009.2014796.
- [20] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. HMM-based text-to-articulatory-movement prediction and analysis of critical articulators. In *Interspeech*, pages 2194–2197, September 2010. URL [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_2194.html](http://www.isca-speech.org/archive/interspeech_2010/i10_2194.html).
- [21] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. An analysis of HMM-based prediction of articulatory movements. *Speech Communication*, 52(10):834–846, October 2010. doi:10.1016/j.specom.2010.06.006.
- [22] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):207–219, January 2013. doi:10.1109/tasl.2012.2215600.
- [23] John E. Lloyd, Ian Stavness, and Sidney Fels. ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In *Studies in Mechanobiology, Tissue Engineering and Biomaterials*, pages 355–394. Springer, 2012. doi:10.1007/8415\_2012\_126.

- [24] Vikramjit Mitra, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1913–1924, September 2011. doi:10.1109/TASL.2010.2103058.
- [25] Kevin G. Munhall, Eric Vatikiotis-Bateson, and Yoh'ichi Tohkura. X-ray film database for speech research. *Journal of the Acoustical Society of America*, 98(2):1222–1224, August 1995. doi:10.1121/1.413621.
- [26] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, and Michael Proctor. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *Journal of the Acoustical Society of America*, 136(3):1307–1311, September 2014. doi:10.1121/1.4890284.
- [27] Aaron Niebergall, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm. Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, 69(2):477–485, April 2012. doi:10.1002/mrm.24276.
- [28] Korin Richmond and Steve Renals. Ultrax: An animated midsagittal vocal tract display for speech therapy. In *Interspeech*, pages 74–77, September 2012. URL [http://www.isca-speech.org/archive/interspeech\\_2012/i12\\_0074.html](http://www.isca-speech.org/archive/interspeech_2012/i12_0074.html).
- [29] Korin Richmond, Phil Hoole, and Simon King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Interspeech*, pages 1505–1508, August 2011. URL [http://www.isca-speech.org/archive/interspeech\\_2011/i11\\_1505.html](http://www.isca-speech.org/archive/interspeech_2011/i11_1505.html).
- [30] Korin Richmond, Zhen-Hua Ling, and Junichi Yamagishi. The use of articulatory movement data in speech synthesis applications: An overview. *Acoustical Science and Technology*, 36(6):467–477, November 2015. doi:10.1250/ast.36.467.
- [31] Dietmar Schabus, Michael Pucher, and Gregor Hofer. Joint audiovisual hidden semi-Markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347, April 2014. doi:10.1109/jstsp.2013.2281036.
- [32] Paul W. Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, May 1987. doi:10.1016/0093-934X(87)90058-7.
- [33] Ingmar Steiner, Korin Richmond, Ian Marshall, and Calum D. Gray. The magnetic resonance imaging subset of the mngu0 articulatory corpus. *Journal of the Acoustical Society of America*, 131(2):EL106–EL111, February 2012. doi:10.1121/1.3675459.

- [34] Massimo Stella, Paolo Bernardini, Francesco Sigona, Antonio Stella, Mirko Grimaldi, and Barbara Gili Fivela. Numerical instabilities and three-dimensional electromagnetic articulography. *Journal of the Acoustical Society of America*, 132(6):3941, December 2012. doi:10.1121/1.4763549.
- [35] Maureen Stone. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, 19(6-7):455–501, January 2005. doi:10.1080/02699200500113558.
- [36] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, July 2012. doi:10.2312/SCA/SCA12/275-284.
- [37] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 1315–1318, June 2000. doi:10.1109/ICASSP.2000.861820.
- [38] John R. Westbury. X-ray microbeam speech production database user’s handbook. Technical report, University of Wisconsin, June 1994. URL [http://www.haskins.yale.edu/staff/gafos\\_downloads/ubdbman.pdf](http://www.haskins.yale.edu/staff/gafos_downloads/ubdbman.pdf).
- [39] Alan A. Wrench and Peter Balch. Towards a 3D tongue model for parameterising ultrasound data. In *18th International Congress of Phonetic Sciences (ICPhS)*, August 2015. URL <http://www.icphs2015.info/pdfs/Papers/ICPHS0768.pdf>.
- [40] Kele Xu, Yin Yang, A. Jaumard-Hakoun, C. Leboulenger, G. Dreyfus, P. Roussel, M. Stone, and B. Denby. Development of a 3D tongue motion visualization platform based on ultrasound image sequences. In *18th International Congress of Phonetic Sciences (ICPhS)*, August 2015. URL <http://www.icphs2015.info/pdfs/Papers/ICPHS0360.pdf>.
- [41] Shuji Yokomizo, Takashi Nose, and Takao Kobayashi. Evaluation of prosodic contextual factors for HMM-based speech synthesis. In *Interspeech*, pages 430–433, September 2010. URL [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_0430.html](http://www.isca-speech.org/archive/interspeech_2010/i10_0430.html).
- [42] Jun Yu, Chen Jiang, and Zengfu Wang. Creating and simulating a realistic physiological tongue model for speech production. *Multimedia Tools and Applications*, pages 1–17, September 2016. doi:10.1007/s11042-016-3929-6.
- [43] Heiga Zen and Tomoki Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Interspeech*, September 2005. URL [http://www.isca-speech.org/archive/interspeech\\_2005/i05\\_0093.html](http://www.isca-speech.org/archive/interspeech_2005/i05_0093.html).
- [44] Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, November 2009. doi:10.1016/j.specom.2009.04.004.

- [45] Yanli Zheng, Mark Hasegawa-Johnson, and Shamala Pizza. Analysis of the three-dimensional tongue shape using a three-index factor analysis model. *Journal of the Acoustical Society of America*, 113(1):478–486, January 2003. doi:10.1121/1.1520538.